ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/228852461

Louis Leon Thurstone in Monte Carlo: creating error bars for the method of paired comparison

Article *in* Proceedings of SPIE - The International Society for Optical Engineering · December 2003

DOI: 10.1117/12.526215

CITATIONS
15

READS
416

1 author:

Ethan Montag
FluxData, Inc.
45 PUBLICATIONS **886** CITATIONS

SEE PROFILE

All content following this page was uploaded by Ethan Montag on 19 December 2013.

The user has requested enhancement of the downloaded file.

Louis Leon Thurstone in Monte Carlo: Creating Error Bars for the Method of Paired Comparison

Ethan D. Montag*
Munsell Color Science Laboratory, Chester F. Carlson Center for Imaging Science
Rochester Institute of Technology, 54 Lomb Memorial Dr., Rochester, NY 14623

ABSTRACT

The method of paired comparison is often used in experiments where perceptual scale values for a collection of stimuli are desired, such as in experiments analyzing image quality. Thurstone's Case V of his Law of Comparative Judgments is often used as the basis for analyzing data produced in paired comparison experiments. However, methods for determining confidence intervals and critical distances for significant differences based on Thurstone's Law have been elusive leading some to abandon the simple analysis provided by Thurstone's formulation. In order to provide insight into this problem of determining error, Monte Carlo simulations of paired comparison experiments were performed based on the assumptions of uniformly normal, independent, and uncorrelated responses from stimulus pair presentations. The results from these multiple simulations show that the variation in the distribution of experimental results of paired comparison experiments can be well predicted as a function of stimulus number and the number of observations. Using these results, confidence intervals and critical values for comparisons can be made using traditional statistical methods. In addition the results from simulations can be used to analyze goodness-of-fit techniques.

Keywords: Psychophysics, Thurstone's Law of Comparative Judgments, paired comparison

1. INTRODUCTION

In order to evaluate the effectiveness of different techniques and algorithms on image quality or image preference psychophysical experiments must be employed to assign numerical scales to the subjective psychological response elicited by the stimuli. Many methods exist for creating scales of sensation and perception (see, for example, Bartleson, Engledrum, Gescheider). Common techniques, such as the classical psychophysical threshold methods or ratio scaling methods can be used to create scales of sensory magnitude. For example, magnitude estimation can be used to construct a ratio scale of perceived lightness that follows Stevens' Law.

However, for judgments of image quality or image preference, these techniques may not be applicable. For example, if one wants to determine which of several printers produces prints that are of highest image quality, there is no one continuous physical parameter that is being manipulated. In addition, the scale that is to be created is not a ratio scale because there exists no concept of "zero quality„. What is desired is a scale that assigns numbers to the psychological percept that allows comparisons of the different stimuli. Such techniques for creating interval scales include different ranking, sorting, and comparison procedures and statistical analyses that transform the subjects' ratings into interval scales.

Of these various techniques, paired comparison has become a popular tool in creating interval scales because of its increased rigor in adhering to the theory and assumptions upon which the transformation of the judgment to a scale is made. Specifically, reference is made to Thurstone's Law of Comparative Judgment.[4] In its original formulation, Thurstone presents a simple theory of the discriminal process and how its nature allows the construction of an interval scale based on comparisons of pairs of stimuli. This has been the starting point for much research in the application and analysis of paired comparison data.[5,6] However, in its original formulation,[4,7] and simplifying assumptions, the analysis of paired comparison data is computationally easy and straightforward.

This said, we have had some problems in implementing Thurstone's Law because the ability to compute confidence intervals is missing in the formulation. Later research has used modifications of Thurstone's Law to produce error

* montag@cis.rit.edu; phone 1 585 475-5096; fax 1 585 475-4444; www.cis.rit.edu/research/mcsl/

Image Quality and System Performance, edited by Yoichi Miyake,
D. René Rasmussen, Proc. of SPIE-IS&T Electronic Imaging,
SPIE Vol. 5294 © 2004 SPIE and IS&T · 0277-786X/04/$15.00

metrics, but this seems to be at a cost of losing the inherent simplicity of Thurstone's original formulation. In this paper we use Monte Carlo simulations to estimate an empirical formula for constructing error bars based on Thurston's Law. It is recognized that a thorough understanding of the underpinnings of statistical distributions, experimental noise, and mathematical precision would lead to a much more satisfactory answer, but for now we are more interested in creating a tool for the analysis and determination of the significance of our data.

2. THURSTONE'S LAW OF COMPARATIVE JUDGMENT

Thurstone[4,7] presents the Law of Comparative Judgment based on the following propositions:

1. Each stimulus gives rise to a discriminal process, which has some value on the psychological continuum of interest.
2. Due to momentary fluctuations (which can be considered as internal fluctuations occurring within or between observers), the value of a stimulus may be higher or lower on repeated presentations. The distribution of this fluctuation can be characterized by a postulated normal distribution (or some other known distribution).
3. The mean and standard deviation of the distribution associated with a stimulus are its internal scale values and discriminal dispersion, respectively.
4. Therefore the distribution of the difference between two stimuli is also normally distributed and it is a function of the proportion that one stimulus is chosen as greater than the other.
5. The difference in scale values, R, between two stimuli, i and j, is:

$$R_i - R_j = z_{ij}\sqrt{\sigma_i^2 + \sigma_j^2 - 2r_{ij}\sigma_i\sigma_j} \tag{1}$$

where R_i and R_j represent the scale values of stimuli i and j, σ_i and σ_j, are the standard deviations of the respective discriminal dispersions, r_{ij} is the correlation between the two discriminal processes, and z_{ij} is the normal deviate (the z-score) corresponding to the proportion of times stimulus j is judged is judged greater along the psychological continuum than stimulus i.

Certain assumptions can be made that result in a simplification of this equation that also lead to a very simple procedure for analysis of paired comparison data. These assumptions, Thurstone's Case V, are:

1. The evaluation of one stimulus along the continuum does not influence the evaluation of the other in the paired comparison ($r_{ij} = 0$).
2. The dispersions are equal for all stimuli ($\sigma_i = \sigma_j$).

These assumptions lead to this formulation of the Law:

$$R_i - R_j = z_{ij}\sigma\sqrt{2} \tag{2}$$

Procedures for testing these assumptions (and the normality of the discriminal distribution) have been provided but for our purposes in paired comparison along continua such as image quality or preference we accept these assumptions. (If we suspected that the change in the discriminal dispersion was affected by the stimulus magnitude, as in Fechner's Law, Thursone's Case III, where ($r_{ij} = 0$), may be applied.)

The analysis of data, according to this equation is as follows:

With n stimuli, n(n-1)/2 stimulus pairs are presented N times for judgment where N observations where N is the product of the number of judgments of each pair made by the total number of observers. A frequency matrix (where the ith column entry is chosen over the jth row entry) of these judgments is constructed and then converted into proportions by diving by N. In turn the proportion are converted into normal deviates and the average of the columns create the interval scale values for the stimuli.

Equation 2 indicates that the standard deviation $\sigma = 1/\sqrt{2}$. We have in the past used this to create confidence intervals using $95\%\ CI = R \pm 1.96(0.707/\sqrt{N})$ where the term $(0.707/\sqrt{N})$ reflects an estimate of the standard error but we have

never quite been satisfied with this because we surmised that the number of stimuli must also play a role in determining the error. Fortunately, the confidence intervals that result from this equation are quite conservative so that our previous experimental findings are not put into jeopardy.

There seems to be no simple solution for determining the confidence intervals based on the Case V solution. Bock and Jones present a generalized version of Case V in which the standard deviation of the discriminal process changes as a function of the scale value (as in Weber's Law). In this formulation the solution as presented above for Case 5 is a preliminary solution that then is weighted using the minimum normit χ^2 analysis.[6]

Alternatively, other distributions, rather than the normal distribution can be used in Case V analysis. Engledrum[2] presents a table of these. Bock and Jones[6] give example of analyses using the arcsine transformation, which provides a simple form for determining confidence intervals:

$$CI = R \pm z_{1-(\alpha/2)}\sqrt{2/(nN)} \tag{3}$$

where $z_{1-(\alpha/2)}$ is the normal deviate corresponding to the chosen α-level (e.g. 1-.05/2 for 95% confidence level). The form of this equation matched our intuition because of the inclusion of the number of stimuli in the experiment but as we will see, it doesn't work for the normal distribution. David[8] presents other methods for determining the statistical significance of scale value differences but these too rely on different analyses of the data.

3. MONTE CARLO SIMULATION

To determine how scale values can vary and then empirically determine the standard deviation of these values for the construction of confidence intervals and tests of significance a Monte Carlo simulation of the paired comparison experiment was performed. By assuming an underlying psychological continuum that conforms to Thurstone's Case V and randomly sampling from normal distributions along this continuum we can recreate the distribution of results that are expected over many repetitions of the experiment.

3.1. The Simulation

For n stimuli, we chose n values as the means of normal distributions all with equal standard deviations. Samples were chosen from each of the n(n-1)/2 pairs of distributions and compared and the results were tallied. This was done N times for the N observations in the experiment. This would constitute one experiment in which the scale values were calculated as described above. This experiment was then repeated many times so that the means and standard deviations of the scale values could be calculated. The mean of the standard deviations of each of the scale values was then calculated to determine the overall standard deviation expected from an experiment with n stimuli and N observations.

So, for example, for n = 6 stimuli, 6 normal distributions with means located at [5 6 7 8 9 10] and a standard deviation of 5 were used for sampling of the stimulus pairs. (A large standard deviation along the psychological continuum was chosen to try and reduce the number of experimental results that contained unanimous judgments.) This was repeated for, say, N = 30 observations. For each pair of n and N, the experiment was repeated 10,000 times and the means and standard deviations of the 6 scale values were calculated. The mean of the 6 standard deviations was taken as the overall observed standard deviation seen in an experiment of that particular size.

The number of stimuli used in the experiment was n = [4 5 6 7 8 10 12 15]. The number of observation was N = [10 20 30 40 50 60]. Each experiment was repeated 10,000 times except when n = 15 which was repeated 5,000 times because of limitations in computer memory. These simulations were repeated twice with the standard deviation of the underlying psychological distribution at a value of 5 and 6. The locations of the means of the distributions along the psychological continuum and their standard deviations did not have an effect on the results, as would be expected from the theory.

Figure 1 shows the results of the simulation. The observed standard deviation is a function of both the number of stimuli and the number of observations.

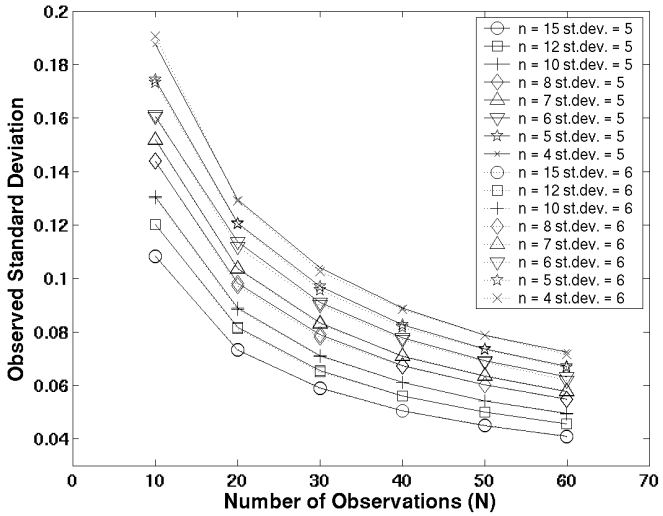
Observed Standard Deviation
Number of Observations (N)
n = 15 st.dev. = 5
n = 12 st.dev. = 5
n = 10 st.dev. = 5
n = 8 st.dev. = 5
n = 7 st.dev. = 5
n = 6 st.dev. = 5
n = 5 st.dev. = 5
n = 4 st.dev. = 5
n = 15 st.dev. = 6
n = 12 st.dev. = 6
n = 10 st.dev. = 6
n = 8 st.dev. = 6
n = 7 st.dev. = 6
n = 6 st.dev. = 6
n = 5 st.dev. = 6
n = 4 st.dev. = 6

Fig.1. Results of the multiple simulations of paired comparison experiments showing the observed average standard deviation of the scale values as a function of the number of observations (N) and stimuli (n).

3.2. Estimates of the observed standard deviation

The next task was to determine a function that characterized the results shown in Fig. 1. For a first try we used the right hand side of Eq. 3, $(z(2/nN)^{0.5})$, letting the value of z vary to find the best fit. The results of this fit are shown in Figure 2. The nonlinear fit of this equation returned a value of 0.845 for z, which corresponds to an α-level of 0.398 or a 0.60 confidence interval. This value does not correspond with the expected confidence interval of approximately 0.84 for 1 standard deviation under the cumulative normal distribution. The percent rms error for this fit is 3.64.

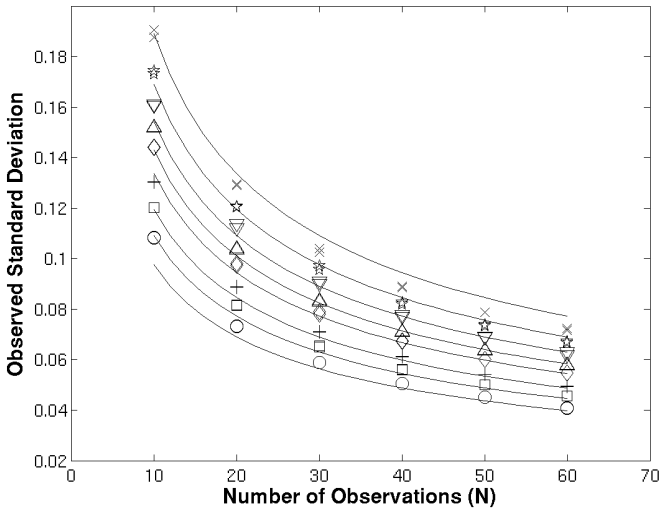

Figure 2. Best fit of the arcsine confidence interval equation to the simulation data.

This fit captures the characteristics of the curves but the standard deviation is underestimated when N is small and n is big and is overestimated when N is big and n is small.

Abandoning an analytic approach a number of equations were fit to the data to find an empirical equation that could capture the observed standard deviation as a function of the number of stimuli and observations. The following equation gave a good fit:

$$\sigma_{obs} = b_1\left(n - b_2\right)^{b_3}\left(N - b_4\right)^{b_5} \tag{4}$$

with $b_1 = 1.76$, $b_2 = -3.08$, $b_3 = -0.613$, $b_4 = 2.55$, and $b_5 = -0.491$. The fit of this equation to the data is shown in Figure 3. The percent rms of this fit was 0.87. The fit is very good, which is not remarkable given that there are 5 parameters. It does seem to overestimate the standard deviation at N = 10. This is not a bad feature considering that with such a small number of observers there is much less power in the experiment and the chance of unanimous judgments occurring by chance is increased.

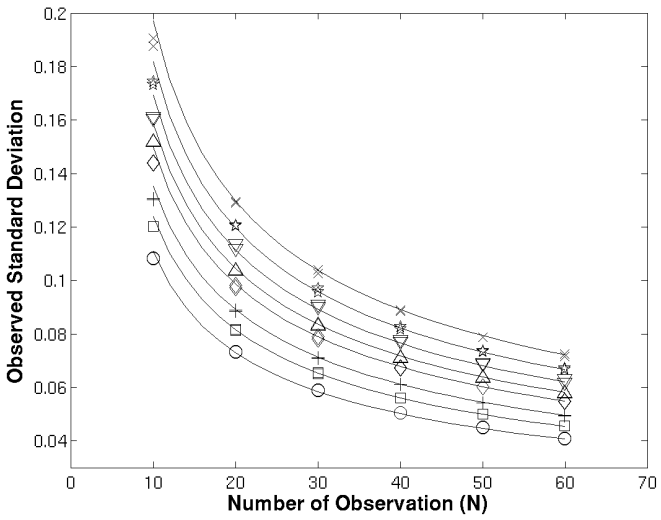

Figure 3. An empirical fit to the simulation data using Eq. (4)

3.3. Verification

Given the empirical estimate of the standard deviation in Eq. 4, we can determine whether using this formulation leads to the predicted results for a new simulation. Arbitrarily selecting n = 5 and N = 33 let us choose mean values on a psychological continuum of [3 4 5 6 7] with a standard deviation of 7. Eq. 4 predicts that the standard deviation is 0.0910. The 95% confidence interval would correspond to ±1.96(0.0910) or ±0.178. The results of 10,000 repetitions of this experiment give an average standard deviation of 0.0906, which is different from the estimate by 0.44%. An average of 489 values (4.89%) were outside the confidence interval defined above which is close to the desired 95% confidence interval. Using the observed average standard deviation, 4.99% of the values were outside the confidence interval. Repeating this for n = 9 and N = 25, the predicted standard deviation is 0.0826. The observed average standard deviation is 0.0815 (a difference of 1.3%) and 4.67% of the values fell outside the confidence intervals. However 5.1% of the values were outside the confidence interval if the average observed standard deviation is used. Based on these two observations, it seems that using Eq. (4) to calculate confidence intervals leads to a slight overestimation of the regions boundaries. Approximately 0.104 units separate the scale values of the results. This indicates that, for example, stimulus 3 is not significantly different than stimuli 2 and 4, but it is significantly different than stimuli 1 and 2.

An analysis of the distributions of the scale values that result from the multiple experimental runs may reveal why these differences may exist for such large simulations. Figure 4 shows the histogram of the scale values for the 3rd item in the first verification example above. This figure also shows the location of the confidence limits calculated above. Notice that the distribution has an overall "normal,, shape for its envelope but it is not smooth. In fact, a Lilliefors test for goodness of fit to a normal distribution rejects the null hypothesis that the distribution is normally distributed. For

comparison, Figure 5 shows a histogram of 10,000 values based on a normal distribution with the observed mean and the computed standard deviation and of the item shown in Fig. 3 along with the 95% confidence limits.
Because the resulting distributions of the scale values from the simulations are not normal, it is likely that the estimates of confidence intervals based on an empirical equation will be incorrect despite the fact that it predicts the computed standard deviation quite well.

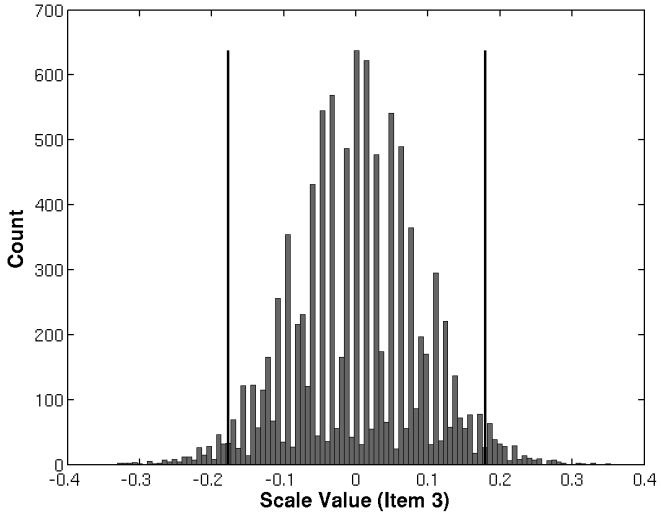

Figure 4. The histogram of the scale values for one of the stimuli in the simulation.

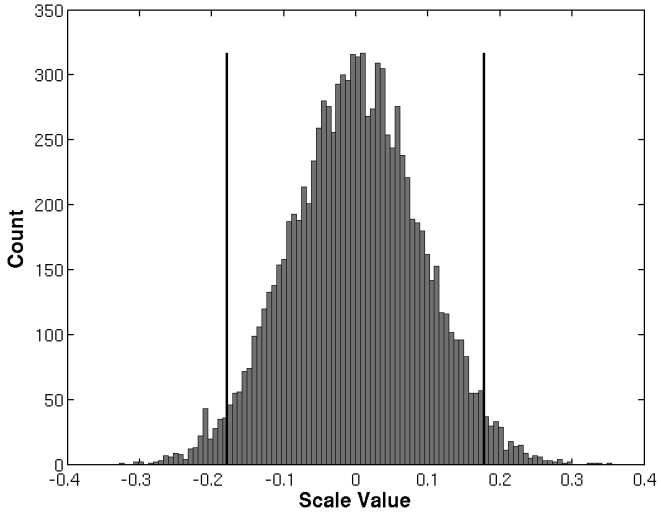

Figure 5. Histogram of 10,000 samples drawn from a normal distribution with the same mean and computed standard deviation as the item in Fig. 4.

3.4. Goodness of Fit

In the course of doing the simulations, Mosteller's χ^2 test of goodness-of-fit was performed on each of the individual experiments in the simulation. This test compares the arcsine transform of the observed proportions to those would be predicted given scale values.[2] This transform is generally valid for large values of N. Figure 6 shows the proportion of

the experiments that were rejected based on this test for the different combinations of numbers of subjects and observations. Surprisingly, this proportion is never less than 0.05 even thought the underlying distributions from which the observations were drawn were normally distributed, conformed to the Case V assumptions, and unidimensional. This would lend one to think that this test is conservative however Engledrum cites Mosteller's reports that the model often appears better than it really is.[2]

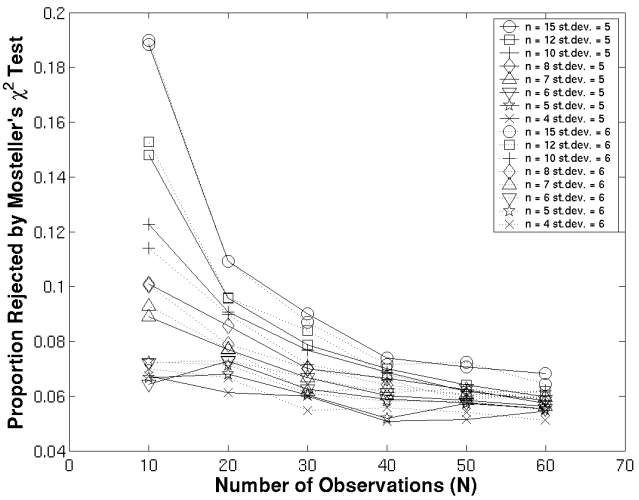
Proportion Rejected by Mosteller's χ^2 Test
Number of Observations (N)
n = 15 st.dev. = 5
n = 12 st.dev. = 5
n = 10 st.dev. = 5
n = 8 st.dev. = 5
n = 7 st.dev. = 5
n = 6 st.dev. = 5
n = 5 st.dev. = 5
n = 4 st.dev. = 5
n = 15 st.dev. = 6
n = 12 st.dev. = 6
n = 10 st.dev. = 6
n = 8 st.dev. = 6
n = 7 st.dev. = 6
n = 6 st.dev. = 6
n = 5 st.dev. = 6
n = 4 st.dev. = 6

Figure 6. Proportions of simulated experiments that were rejected by Mosteller's χ^2 test of goodness-of-fit.

For the verification example above with 5 stimuli, one of the experiments that was evaluated as having a bad fit was examined more closely to determine why it was rejected by Mosteller's test. Table I shows the proportion matrix for this experiment. Table II shows the proportion matrix based on the results of the experiment. Figure 7 shows the results of the paired comparisons for this simulation. Part A represents the underlying means on the psychological continuum so that the winner of each pair has an arrow pointing to the loser. Part B shows the results from the simulated experiment. By chance, stimulus 1 was chosen over stimulus 5 (open arrow) creating a *circular triad*[2] indicating a logical inconsistency. Part C shows the paired comparisons predicted by the results of this experiment's analysis. By chance, the occurrence of circular triads causes results that will fail the goodness-of-test. The simulations therefore give us an indication of the probability of the chance occurrence of data that fail the goodness-of-fit test even when all the assumptions of case V are obeyed.

p-matrix	Stim. 1	Stim. 2	Stim. 3	Stim.s 4	Stim. 5
Stim. 1	0.50	0.73	0.73	0.58	0.45
Stim. 2	0.27	0.50	0.76	0.55	0.56
Stim. 3	0.27	0.24	0.50	0.55	0.64
Stim. 4	0.42	0.45	0.45	0.50	0.58
Stim. 5	0.55	0.42	0.36	0.42	0.50

Table I: Observed matrix of proportions for rejected experiment.

p-matrix	Stim. 1	Stim. 2	Stim. 3	Stim. 4	Stim. 5
Stim. 1	0.50	0.57	0.66	0.62	0.65
Stim. 2	0.43	0.50	0.60	0.55	0.58
Stim. 3	0.34	0.40	0.50	0.45	0.48
Stim. 4	0.38	0.45	0.55	0.50	0.53
Stim. 5	0.35	0.42	0.52	0.47	0.50

Table II: Predicted matrix of proportions from the results of the same experiment as inTable I.

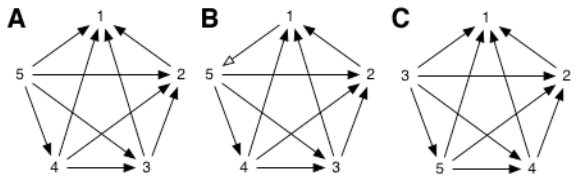

Figure 7. Paired comparisons based on: A) the psychological continuum,
B) the results of the simulated experiment, and C) the predicted results of the experiment.

3.5. Implementation

Despite the problems observed above, the use of an empirical equation may be helpful for practical application in the analysis of paired comparison data. Further simulation can be done to better refine the choice of values for a chosen confidence region. In the absence of a more refined solution, these results can still be used as a tool to help us analyze our data.

3.5.1. Paired Comparisons of Scale Values

Besides its use for determining confidence intervals, the statistics determined via simulation can be useful in determining more stringent confidence intervals for multiple comparisons of scale values. It is recognized that multiple comparisons of pairs of means in the analysis of variance leads to an increase in errors. It must also be true for the comparison of multiple scale values in paired comparison experiments. Keppel[9] enumerates a number of planned and post hoc tests such as the Scheffé, Dunnett, and Tukey tests for controlling error rate for multiple comparisons. Using an empirical value of the standard deviation in the place of the standard error in these tests may prove useful.

3.5.2. Comparing Experiments and Methods of Reducing Labor

Although the comparison of results from separate experiments should be avoided, in practice one may desire such a comparison. Because of the normalization of scale values in the analysis of paired comparison, at least two stimuli need to common to experiments in order to account for the multiplicative and additive shifts that result in interval scales produced in different experiments. Likewise, the confidence intervals must be scaled appropriately to compare experiments. Taking into account that the experiments may have different numbers of stimuli and observers, the confidence intervals in the combined results will be different and must be also scaled appropriately.

Likewise, for experiments with large numbers of stimuli, methods of reducing labor have been devised[5] which smaller sets of comparisons are done that only sample a subset of all the possible comparisons between stimuli. In these cases the number of observers is constant but these subsets may have different numbers of stimuli. Therefore in the analysis of these data, the confidence intervals may be different when these subsets are combined with the appropriate multiplicative and additive shifts taken into account.

4. CONCLUSIONS

After recognizing that analysis of paired comparison data using Case V of Thurstone's Law of Comparative Judgment is a straightforward and simple process, we faced the problem that a simple expression for the variability of the results seems to lacking. A Monte Carlo simulation of paired comparison experiments was performed in order to empirically determine the standard deviation of experiments of various sizes and it was found that the variability in the experiments is a function of both the number of stimuli and the number of observations.

Based on these simulations, an empirical formula for the standard deviation was fit to the observed values. This equation fit the data well because the data from the simulations was rather well behaved; it is no surprise that an equation with 5 parameters could fit a smooth monotonic curve. No claim is made that this equation is the best fit possible considering the multitude forms of equations that can be fit. In addition it is recognized that the form of this equation does not have any theoretical significance.

In verifying the use of this equation, it appears that it may slightly overestimate the 95% confidence intervals so a somewhat more stringent criterion may be used to account for this. Further simulations can lead to insight into this problem and can be used to more directly determine confidence intervals rather than the standard deviation. This might be a good course for further research because it was observed that the distributions of the scale values from the simulation were not normally distributed. Each experiment in the simulation was evaluated for goodness-of-fit by Mosteller's χ^2 test. These results indicated conservative performance of this test given the experimental assumptions.

It was pointed out that when multiple scale values from paired comparison experiments are compared, there could be an increase in error rate. It was proposed that the standard deviation derived empirically from the simulations might have utility in correcting these error rates by application in known statistical tests for multiple comparisons of means.

Special care must be used in comparing results from different experiments or using methods for reducing labor in large paired comparison experiments. The normalization of the scales in separate experiments or subsets of larger experiments require multiplicative and additive correction of the data. In addition, the differences in the number of stimuli and observations in the different data sets will require scaling of confidence intervals that may be of different size.

REFERENCES

1. C. James Bartleson and Franc Grum (Eds.), *Optical Radiation Measurements*, Vol 5. Academic Press, New York, NY, 1984.
2. Peter G. Engeldrum, *Psychometric Scaling*, Imcotek Press, Winchester, MA, 2000.
3. George A. Gescheider, *Psychophysics: The Fundamentals*, Lawrence Erlbaum Assoc., Publishers, Mahwah, NJ, 1997.
4. Leon Lewis Thurstone, "Psychophysical Analysis,,, *Am. J. Psych.* 38:368-389, 1927.
5. Warren S. Torgerson, *Theory and Methods of Scaling*, J. Wiley & Sons, New York, 1958.
6. R. Darrell Bock and Lyle V. Jones, *The Measurement and Prediction of Judgment and Choice*, Holden-Day, San Francisco, CA, 1968.
7. Leon Lewis Thurstone, "A Law of Comparative Judgment,,, *Psychological Review*, 34:273-286, 1927.
8. H. A. David, *The Method of Paired Comparison*, Hafner Press, New York, NY, 1969.
9. Keppel, Geoffrey, *Design and Analysis A Researcher's Handbook, 2nd Ed.*, Chapter 12, Prentice-Hall, Inc., Englewood Cliffs, NJ, 1982.

View publication stats